CEUR-WS.org/Vol-2517/T2-10.pdf

KCE_DALab-APDA@FIRE2019: Author Profiling and Deception Detection in Arabic using Weighted Embedding

Sharmila Devi V[1], Kannimuthu S[1], Ravikumar G[2], and Anand Kumar M[3]

[1] Department of Information Technology, Karpagam College of Engineering, Coimbatore
[2] Department of Computer Science and Engineering, CIET, Coimbatore
sharmiladevi1002@gmail.com
[3] Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, India
m_anandkumar@nitk.edu.in

Abstract. This paper explaining the work submitted on Author Profiling and Deception Detection in Arabic Tweets shared task organized at the Forum for Information Retrieval Evaluation (FIRE) 2019. The first task Author profiling illustrates identifying the categories of authors based on the Arabic tweets. In the second task, the aim is to Detect deception in Arabic for two genres such as Twitter and News. Deception detection means that the automatic way of identifying false messages in the text content on social network or news. For each task, we have submitted three different systems. For submission 1, we have used the Term Frequency and Inverse Document Frequency (TFIDF) based Support Vector Machine classification and in submission 2, we have used fastText classifier. For submission 3, we have proposed a low dimensional weighted document embedding (TFIDF + Word embedding) with SVM classification. We have attained second place in the Deception detection and third in Author profiling. The performance difference between the top team results and the submitted runs are only 3.34% for Author profiling and 1.16% for Deception detection.

Keywords: Author profiling · Deception detection · Arabic tweets · Machine Learning · TFIDF · Word embeddings · fastText Classifier · Weighted document embeddings.

1 Introduction

In our busy day-to-day life, a computer-based technology, social media plays a major role in sharing of information, ideas, thoughts from one people to another. Most of the people used to send their personal messages, documents, videos and

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India

photos through social media network such as Twitter, Facebook, WhatsApp etc. Author profiling is the method which analyse the demographic features of an author such as age, gender and the language varieties. Some of the applications of author profiling are forensics, security, marketing, etc. For example, in the marketing field, it is useful to find which profile of the customer like or dislike the product. This analysis will help companies for better market segmentation. From a forensic viewpoint, it is important to find out the profile of the person who wrote the suspicious text. Deception detection is the method of analysing whether the given message is lie or truth. The rest of the paper will briefly as follows: In section 2, we discuss the literature survey about the author profiling and deception detection in various languages. Section 3 mentions the data set description and the statistics. In section 4, we explain the methodology and section 5 discusses the results obtained. In section 6, we conclude the paper with limitations and future work.

2 Related Works

The peculiarities of the Arabic dialectal varieties are used in social media and the annotation framework is proposed in [1]. The suspicious message of the author is whether a potential threat or not is focused in Arabic Author Profiling for Cyber-Security project [2]. The framework for improving the deception detection accuracy for online digital news veracity is proposed in [3]. Bayesian classification and K- means clustering algorithm to find out the deception detection in the twitter profile characteristics is proposed to analyze the user behavior [4]. Various features extraction methods proposed in deception detection from Arabic Twitter post [5]. The accuracy gained for the SVM with trigram over other classifiers is 91.55%. Arabic word correction to manipulate the vulnerability is explained in [6]. They achieved accuracy of 96.5% for detecting abusive Arabic tweets.

Author profiling system for Urdu is proposed [8] by word and character-based term frequency and TFIDF features and support vector machine classifier. Weighted embeddings based on a novel median-based loss function is explained [9] with the experimental results on Wikipedia and twitter data. Embedding variations to the doc2vec embedding on a new evaluation task using Trip advisor reviews, and also the CQADupStack benchmark are proposed in [10]. Word mover's embedding to enable the unsupervised document embedding from pre-trained word embeddings is proposed in [11]. Identification of the age and gender form blog authors are proposed [12] and the experiments on information retrieval features yielded best predictions.

3 Dataset Description

The dataset for Arabic author profiling is given as five different categories where each consists of three natives. The details of the nativity are given in the overview of the shared task [14]. The dataset consists of three age groups (25, Between 25 and 34 and Above 35) and two genders (male and female) in all the categories. The primary difference between the given deception and profiling dataset is in the representation. In Author profiling, each XML file which consists of 100 tweets needs to be labeled as gender, age group, and language variety. But in Deception detection, each tweet should be identified whether it is truth or lie. Two different domains such as News and Tweets are given for deception detection. We have submitted 6 runs for Deception detection.

All the five training dataset of author profiling and deception detection are completely balanced and the number of documents in different classes are given on Table 1 and 2.

Table 3 shows the average tokens per line in the deception detection dataset. For the news genre, the average tokens are similar for training as well as testing. Conversely, on Twitter, the average token size in test data is more compared with the train data. Table 4 explains the average token size of the Arabic author profiling for five training dataset.

Table 1. Author Profiling Data Description

Age		Gender		Nativity	
Under	150	Male	225	Native-1	150
Above	150	Female	225	Native-2	150
Between	150			Native-3	150

Table 2. Deception Detection Data Description

Dataset	Train	Test	Train-Truth	Train-Lie
Qatar-News	1443	370	678	765
Qatar-Twit	532	241	259	273

Table 3. Statistics of Deception Detection Dataset

	Dataset	Tokens	Average
Train	Qatar-News	25792	17.87387387
	Qatar-Twitter	10044	18.87969925
Test	Qatar-News	6635	17.93243243
	Qatar-Twitter	4838	20.0746888

Table 4. Statistics of Author Profiling Dataset

Nativity		Num of Files	Num of Sentences	Total Tokens	Average Tokens
DZ AG IQ	TRAIN	450	45000	612625	1361.38889
	TEST	144	14400	196697	1365.95139
KW LBSY LY	TRAIN	450	45000	541062	1202.36
	TEST	144	14400	196320	1363.33333
MA OM PSJO	TRAIN	450	45000	535074	1189.05333
	TEST	144	14400	179956	1249.69444
QA SA SD	TRAIN	450	45000	591206	1313.79111
	TEST	144	14400	181810	1262.56944
TN UAE YE	TRAIN	450	45000	712111	1582.46889
	TEST	144	14400	237461	1649.03472

4 Methodology

We have totally submitted three methods which are based on TFIDF features with SVM classifier, word bi-grams with fastText classifier and TFIDF weighted document embeddings. We have submitted 21 runs for Arabic Author profiling and Deception detection. In the case of deception detection, we have tried the same approaches followed for the Arabic author profiling task. The three methods are explained below.

Submission-1: The first run is based on the conventional method where we have used the word and character n-gram features with SVM classifier [8]. Word uni-grams and character bi-grams, trigrams and four-grams are considered as features. Out of all features, we have considered a maximum of 5000 features for words and 5000 for characters. These feature values are weighted with TFIDF values. The final feature matrix is given to the Linear SVM for classification. The SVM parameters are L2 norm for a penalty with C value 1 and multi-class using one versus rest. We have followed the same method for Arabic author profiling and Deception detection.

Submission-2: In the second run, we have used the well-known fastText embedding and classifier [7] for profiling the Arabic authors and identifying the deception. The fastText classifier is compatible for the sentence classification, task so for Deception detection we have used the fastText classifier as such. But in the case of Author profiling task, the XML file is input. Fortunately, all the training as well as testing XML files are made from equal (100) tweets. So we have modified the input as individual tweets and trained as a sentence classification task. After tagging the tweets during testing, we have counted the labels of each XML file and select the maximum label as a label for the corresponding XML file. The main drawback of this approach is to infer the cross-validation results. The parameters of fastText are fixed as follows, word bi-grams, learning rate lr=0.25 and 40 epochs. We have used softmax as the loss function.

Submission-3: We have developed the weighted word embedding model for the third submission. Here, we have used the Arabic pre-trained word vectors from Arabic tweets and web pages [13]. The complete architecture of the model is shown in Figure 1. In the case of Author profiling, initially word unigram features are vectorized using conventional TFIDF vectorizer. The maximum features are limited to 5000, so each XML document is represented as 5000 unique words. All the XML documents in the training data are TFIDF vectorized with maximum feature size of 5000. The existing skip-gram based Arabic pre-trained vectors [13] of size 300 are used to create the embedding matrix for the unique words. The words which are not present in the pre-trained vectors are considered as unknown words, for these words the embeddings are generated randomly from the word vectors. Finally, we have taken the dot product between the TFIDF and embedding matrix which results in the document transformed to low dimensional document vectors. These set of vectors are considered as TFIDF weighted document embeddings which are further trained using SVM.

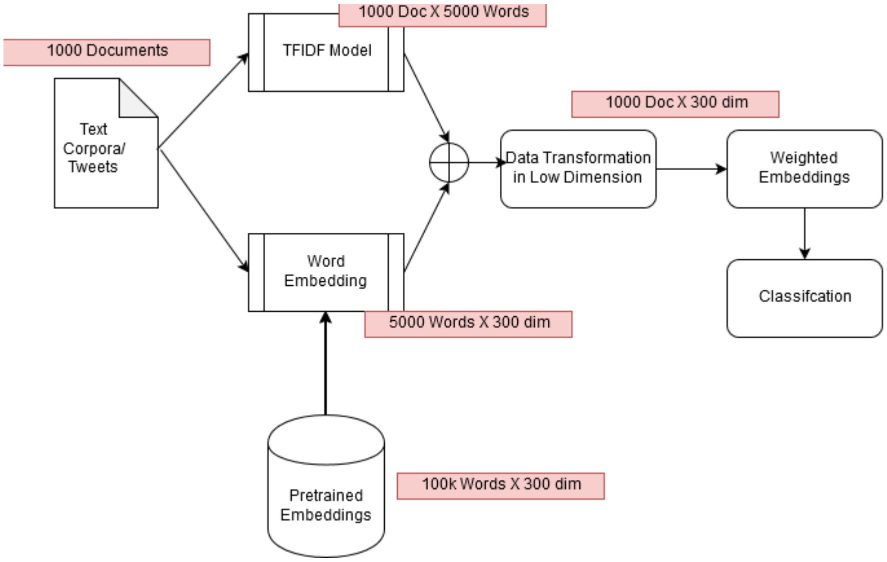

Fig. 1. Weighted Document Embedding Framework.

5 Results

Table 5 and 6 illustrates the top three team's performance on the shared task. Table 7 and 8 shows the Arabic Author profiling and Deception detection accuracies of the three proposed methods. These results are given by the task organizers [14]. We have attained the third position in author profiling and second in Deception detection. The best accuracy of our submissions obtained for

Author profiling in Arabic tweets for gender it is 0.7667, age it is 0.5722 and the variety it is 0.9694. The performance also evaluated jointly where the accuracy gained is 0.4222. The top accuracy gained for Deception detection for news it is 0.7331 and for Twitter, it is 0.8541, the average performance of the accuracy is obtained as 0.7887.

Table 5. Top three team's results of Author Profiling

TEAM-RANK	RANK	TEAM	GENDER	AGE	VARIETY	JOINT
1	1	DBMS-KU.2	0.7944	0.5861	0.9722	0.4556
2	2	Nayel.1	0.8153	0.5708	0.975	0.4486
2	3	Nayel.3	0.8014	0.5792	0.9708	0.4486
1	4	DBMS-KU.3	0.7833	0.5819	0.9778	0.4444
1	5	DBMS-KU.1	0.7778	0.5792	0.9736	0.4347
3	6	KCE_DAlab.sub1	0.7667	0.5722	0.9583	0.4222

Table 6. Top three team's results of Deception Detection

TEAM-RANK	RANK	TEAM	NEWS	TWITTER	AVERAGE
1	1	Nayel.3	0.7542	0.8464	0.8003
1	2	Nayel.1	0.7417	0.8463	0.794
2	3	KCE_DAlab.sub1	0.7232	0.8541	0.7887
2	4	KCE_DAlab.sub2	0.7331	0.8293	0.7812
3	5	DBMS-KU.2	0.7352	0.8125	0.7739

Table 7. Author Profiling Results

Submission	GENDER	AGE	VARIETY	JOINT
KCE_DAlab.sub1	0.7667	0.5722	0.9583	0.4222
KCE_DAlab.sub2	0.7458	0.5708	0.9694	0.4125
KCE_DAlab.sub3	0.7444	0.5028	0.9583	0.3694

6 Conclusion and Future Work

In this paper, we illustrate the work on the identification of age, gender and language variety in author profiling and deception detection in Arabic (APDA). Using the given training dataset, we have developed three systems. We have used the Term Frequency and Inverse Document Frequency and SVM, fastText classifier method and weighted word embedding with SVM. Compared with the

Table 8. Deception Detection Results

Submission	NEWS	TWITTER	AVERAGE
KCE_Dalab.sub1	0.7232	0.8541	0.7887
KCE_Dalab.sub2	0.7331	0.8293	0.7812
KCE_Dalab.sub3	0.6613	0.6791	0.6702

traditional model the most expected weighted embeddings attained less accuracy. The main reason for less accuracy is that the certain words in the given dataset are not present in the pre-trained model. Even though, we have used the pre-trained model of Arabic tweets, around 30% of unknown words present in the training data. This can be resolved with the recent character-specific word embeddings. With this 30% of information loss, the performance of the proposed low-dimensional document embedding on Author profiling attained decent accuracy. In the future, this can be enhanced with character-specific embedding and retrain the pre-trained models.

References

1. Zaghouani, Wajdi, and Anis Charfi. ”Guidelines and Annotation Framework for Arabic Author Profiling.” arXiv preprint arXiv:1808.07678 (2018).
2. Rosso, Paolo, Francisco Rangel, Bilal Ghanem, and Anis Charfi. ”ARAP: Arabic Author Profiling Project for Cyber-Security.” Procesamiento del Lenguaje Natural 61 (2018): 135-138.
3. Eembi@ Jamil, Normala Che, Iskandar Ishak, and Fatimah Sidi. ”Deception detection approach for data veracity in online digital news: Headlines vs contents.” AIP Conference Proceedings. Vol. 1891. No. 1. AIP Publishing, 2017.
4. Alowibdi JS, Buy UA, Philip SY, Ghani S, Mokbel M. Deception detection in Twitter. Social network analysis and mining. 2015 Dec 1;5(1):32.
5. Al-Saif, Hissah, and Hmood Al-Dossari. ”Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques.” International Journal of Advanced Computer Science and Applications 9.10 (2018): 377-387.
6. Abozinadah, Ehab A., and J. H. Jones. ”Improved micro-blog classification for detecting abusive Arabic Twitter accounts.” International Journal of Data Mining and Knowledge Management Process (IJDKP) 6.6 (2016): 17-28.
7. Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. ”Bag of tricks for efficient text classification.” arXiv preprint arXiv:1607.01759 (2016).
8. Sharmila Devi, V., Kannimuthu, S., Ravikumar, G., Anand Kumar, M. ”KCe_Dalab@maponsms-Fire2018: Effective word and character-based features for multilingual author profiling” (2018) CEUR Workshop Proceedings, 2266, pp. 213-222.
9. De Boom, Cedric, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. ”Representation learning for very short texts using weighted word embedding aggregation.” arXiv preprint arXiv:1607.00570 (2016).
10. Schmidt, Craig W. ”Improving a tf-idf weighted document vector embedding.” arXiv preprint arXiv:1902.09875 (2019).

11. Wu, Lingfei, Ian EH Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. ”Word Mover’s Embedding: From Word2Vec to Document Embedding.” arXiv preprint arXiv:1811.01713 (2018)
12. Weren, Edson RD, Anderson U. Kauer, Lucas Mizusaki, Viviane P. Moreira, J. Palazzo M. de Oliveira, and Leandro K. Wives. ”Examining Multiple Features for Author Profiling.” (2014).
13. Abu Bakr Soliman, Kareem Eisa, and Samhaa R. El-Beltagy, AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP, in proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, UAE, 2017.
14. Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., Snchez-Junquera, J.: Overview of the track on author profiling and deception detection in arabic. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15 (2019)